UNIVERSITY COLLEGE DUBLIN

MIS41270 Data Management and Mining

Insurance Data Analysis Project

Supervisor: Aoife D'Arcy

Tao, Baizhou (19211033)

05 April 2021

Assess analytics approaches

In the context that the company is going to launch a new home insurance product in the coming months, the marketing department needs data-driven analysis and corresponding suggestions to maximize the effects of future marketing activities. The first support they need is a customer profile to depict the characteristics of our customers so that the marketing team is able to launch marketing campaigns more effectively. Another support is predictive analytics. Considering the proper marketing channels are able to increase the purchase conversion rate, predicting customers' preferred communication channels is an essential support.

All in all, our team plans to implement two machine learning technologies to achieve the business requirement: unsupervised learning and supervised learning.

1.1 unsupervised learning

Exactly, after consideration, there are 2 potential possibilities of analyzing methods which can be used to make useful strategies for marketing new products.

1) PCA (Principal Component Analysis)

It is a common data analysis method, often used for dimensionality reduction of high-dimensional data and can be used to extract the main feature components of the data.

2) K-means (k-means clustering algorithm)

Clustering is a process of classifying data on certain aspects of data, and clustering techniques are often referred to as there is no monitoring. K-means cluster as the most famous division cluster algorithm, is a technique that discovers such an intrinsic structure because of the simplicity and efficiency.

1.2 supervised learning: classification

In order to achieve higher and more stable prediction performance, we are going to test different algorithms and finally ensemble the ones with better performance together. The possible classification models are as follows: Random forest, XGBoost, Gradientboosting, SVC and other classifiers. Except tuning the hyperparameter of the algorithms, a genetic programming-based tool TPOT (Trang, Weixuan and Jason, 2020) can also help us to generate some good weak estimators. After the cross validation, the final prediction report will be proposed to evaluate the accuracy of each preferred channel.

Data quality report

The first step is to get a snapshot of the data. The data set consists of 4090 rows and 20 columns in total.

Descriptive statistics of data can not only reveal the mean, standard deviation, minimum value, maximum value and other information of numerical data, but also help us find outliers. The distribution information of the population can be seen from the descriptive statistics of the character data.

Counting the number of null values for each column allows people to process data in different ways for various columns. HealthType, HealthDependentsAdults, and HealthDependentsKids have the same number of null values. It should be checked whether uninsured persons did not complete the corresponding columns of HealthDependentsAdults and HealthDependentsKids. Since it is not used as an analysis variable, the null value of the Occupation column can be left alone. The other columns may consider replacing null values with other characters or deleting them directly.

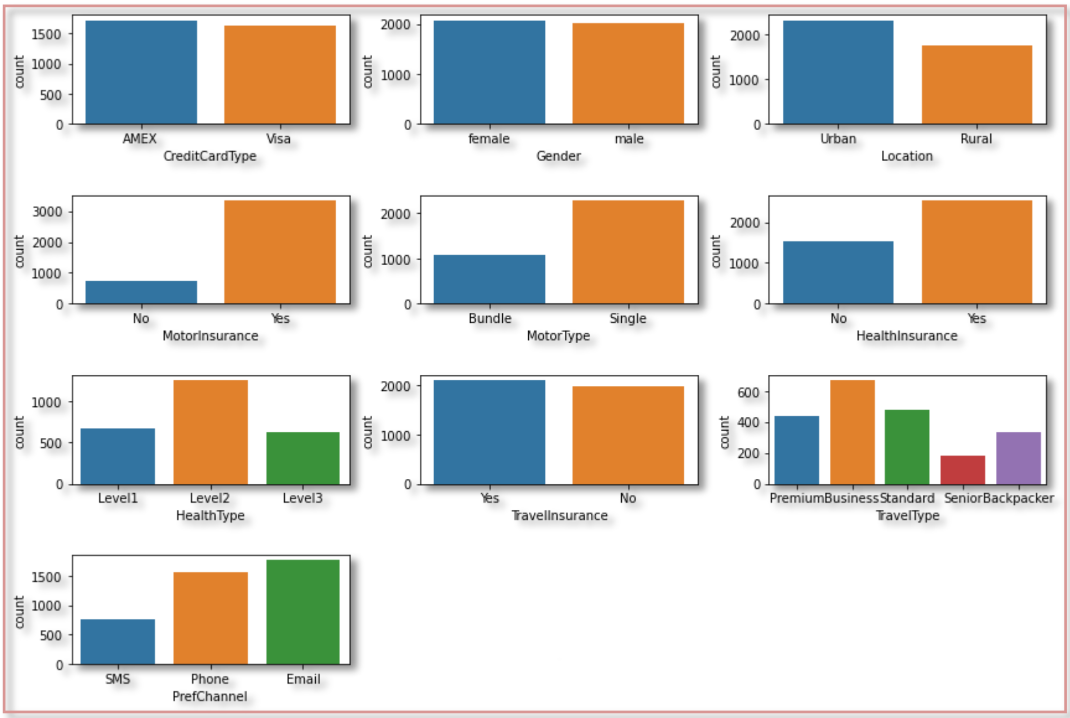
Detailed information on each variable can be found in the following two graphs.

Numerical Features									
	Count	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**	**Miss**
CustomerID	4090	2604.48	1498.31	1	1295.25	2594.5	3908.75	5200	0.00%
Age	4090	41.39	15.99	-44	22	46	50	210	0.00%
MotorValue	3361	23450.91	11985.63	-25686	14837	25045	32289	325940	17.82%
HealthDependentsAdults	2543	0.82	0.65	0	0	1	1	2	37.82%
HealthDependentsKids	2543	1.75	1.11	0	0	2	3	3	0.00%

Categorical Features						
	Count	**Unique**	**Mode**	**Mode Freq.**	**Mode %**	**Miss**
Title	4090	4	Mr.	1950	47.68%	0.00%
GivenName	4090	897	Morgan	27	0.66%	0.00%
MiddleInitial	4090	25	A	459	11.22%	0.00%
Surname	4090	524	Henderson	38	0.93%	0.00%
CreditCardType	3368	2	AMEX	1728	51.31%	17.65%
Occupation	2534	1591	HIV/AIDS nurse	6	0.24%	38.04%
Location	4090	2	Urban	2323	56.80%	0.00%
MotorInsurance	4090	2	Yes	3361	82.18%	0.00%
MotorType	3361	2	Single	2287	68.05%	17.82%
HealthInsurance	4090	2	Yes	2543	62.18%	0.00%

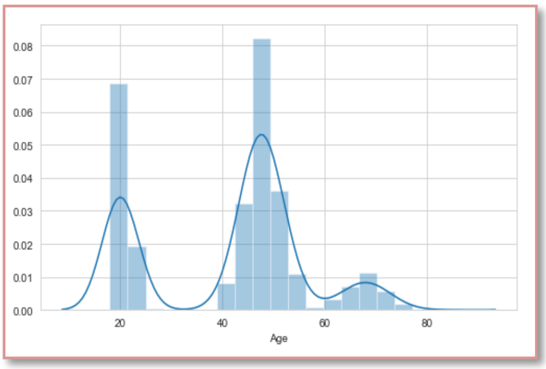
HealthType	2543	3	Level2	1254	49.31%	37.82%
TravelInsurance	4090	2	Yes	2108	51.54%	0.00%
TravelType	2108	5	Business	669	31.74%	48.46%
PrefChannel	4090	3	Email	1768	43.23%	0.00%

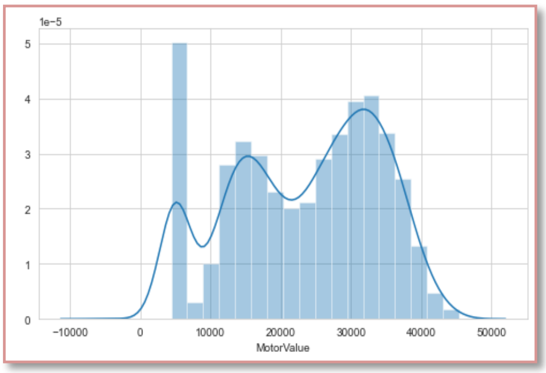
According to the above content, data cleaning was carried out. Please refer to Notebook for the process of data cleaning. After the data cleaning, the data in the dataset are analysed simply. The following figure shows the proportions of different categories in each variable.

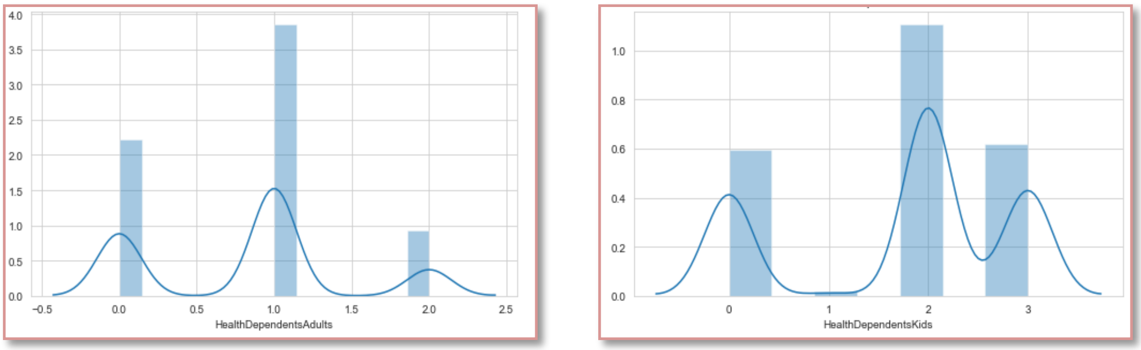


Categorical features histogram

The distribution of several variables can be obtained by statistical analysis of numerical data.

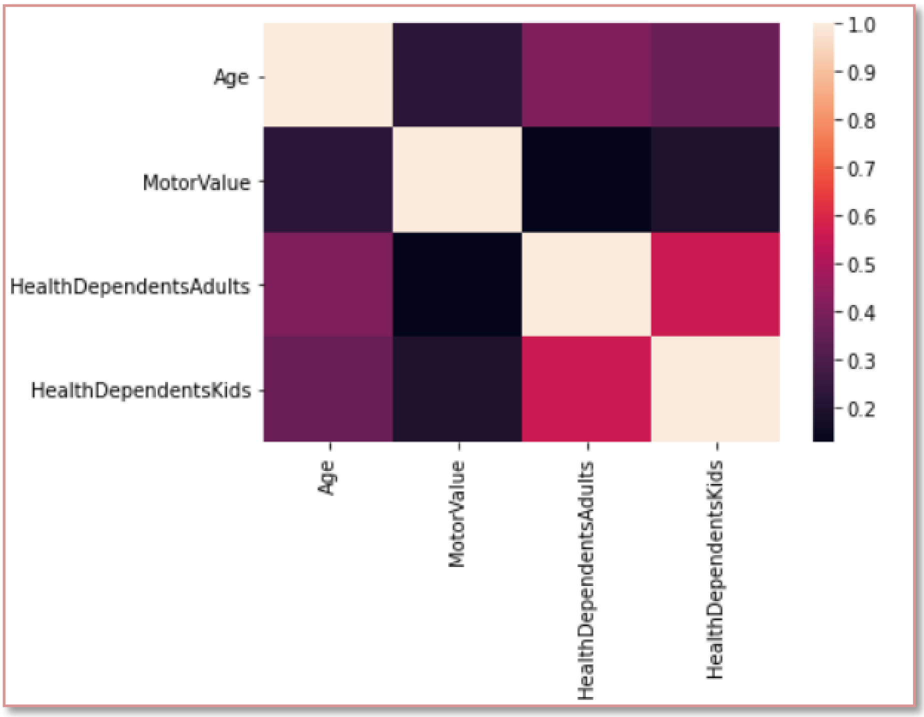






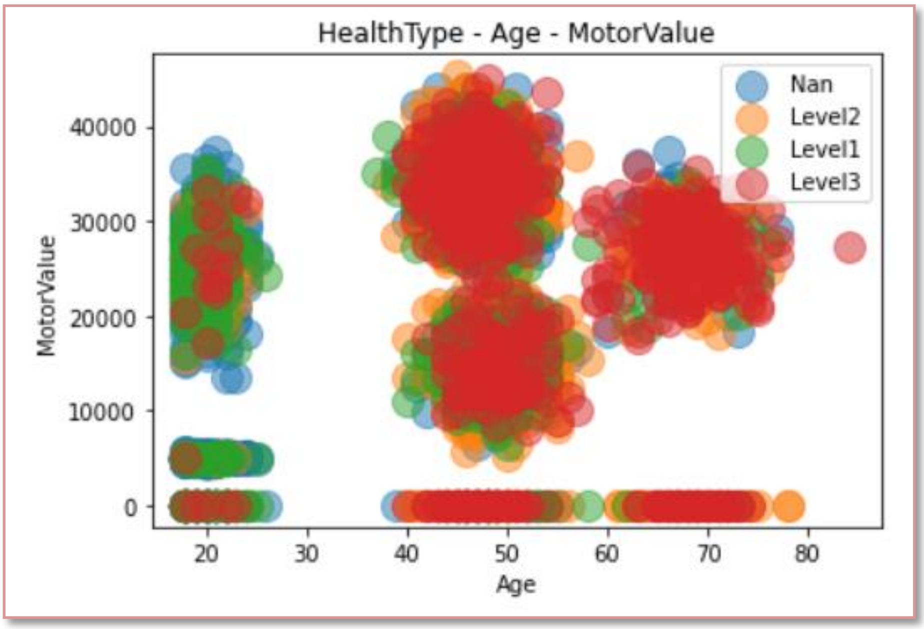
Numerical features histogram

The correlation between the four types of numerical data can be seen in the correlation coefficient heat map below.



Numerical features heatmap

Then the distribution between health types with major numerical data was explored. For example, the figure below shows the distribution of health type with age and MotorValue.



Distribution of health type with age and MotorValue

Customer profile

Based on existing data, unsupervised learning is used to construct Customer portraits. After building the customer portrait, Insure ABC can firstly promote new products HouseInsurance according to the characteristics of different user groups. Then it can also promise the rapid classification of new users in the future. To construct a Customer portrait, we will go as follows.

1.1 Feature Engineering

For a machine learning problem, data and features determine the upper limit of machine learning, and models and algorithms only approach this upper limit.

So, the feature engineering should be first considered and implemented.

Feature engineering, as the name suggests, is a series of engineering processes on raw data, which are refined into features and used as input for algorithms and models. In essence, feature engineering is a process of representing and presenting data. In actual work, feature engineering aims to remove impurities and redundancies in the original data, and design more efficient features to characterize the relationship between the solved problem and the prediction model.

According to the resource, all data is structured data. The structured data type can be regarded as a table of a relational database. Each column has a clear definition, including two basic types: numeric and categorical; each row of data represents information about a sample. So, Feature Engineering can go as follows.

1) Deleting the useless features

According to the data exploration, such useless features can be dropped simply,

Useless features =

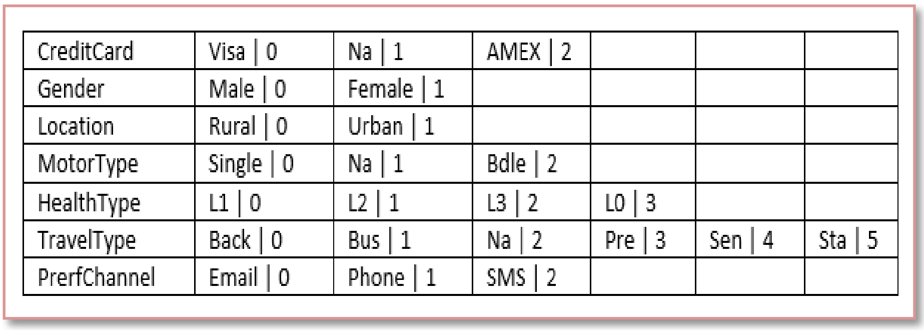
['CustomerID','Title', 'GivenName', 'MiddleInitial', 'Surname', 'Occupation', 'MotorInsurance', 'HealthInsurance', 'TravelInsurance']

Because that ‘CustomerID’ and some other name relative features are just appearing as different labels which provide zero contribution for analyzing. Besides that ‘Occupation’ has too many unique values and the feature is also impossible to be divided by occupation field or job title. As for the other three features, according to the before data exploration, ‘MotorInsurance' can be seen as zero MotorValue and other values of MotorValue; ‘HealthInsurance' can be divided into Level0 and other levels in ‘HealthType’; 'TravelInsurance' is equal to different categories of ‘TravelType’and no travel.

2) Encoding for categorical features

We apply LabelEncoder on categorical data because that categorical data cannot be used straightly in PCA and K-means. And here is another reason to use LabelEncoder, we firstly try to use OneHotEncoder, a widely used feature encoding method, to code the Categorical data, but after PCA-process which keeps 3 features, the explained variance ratio of PCA-Features is so low that means new PCA-features lose too much raw data and reduce the accuracy of next K-means clustering. So, I change another method of Encoding - Label encoding to promise the effectiveness of PCA-process.

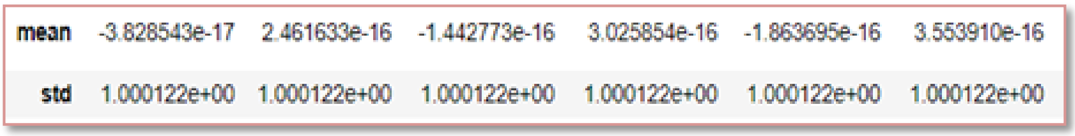
The actual meaning of different codes is shown in the following table.



CreditCard	Visa | 0	Na | 1	AMEX | 2			
Gender	Male | 0	Female | 1				
Location	Rural | 0	Urban | 1				
MotorType	Single | 0	Na | 1	Bdle | 2			
HealthType	L1 | 0	L2 | 1	L3 | 2	L0 | 3		
TravelType	Back | 0	Bus | 1	Na | 2	Pre | 3	Sen | 4	Sta | 5
PrerfChannel	Email | 0	Phone | 1	SMS | 2			

3) Features Normalization

After analyzing the data of different features, diverse feature's std and mean are quite different that means we cannot do dot product calculations for these features. So, we use Normalize to all data to make them mean=0 and std=1. The sample of normalizing results can be seen in the below chart.



mean	-3.828543e-17	2.461633e-16	-1.442773e-16	3.025854e-16	-1.863695e-16	3.553910e-16
std	1.000122e+00	1.000122e+00	1.000122e+00	1.000122e+00	1.000122e+00	1.000122e+00

4) PCA

There are 11 features in the data, it cannot be used to show a visualization, and different features perhaps have relationships between each other. Exactly, PCA is a common data analysis method, often used for dimensionality reduction of high-dimensional data and can be used to extract the main feature components of the data. It also has such significant advantages:

a) Make the data set easier to use.
b) Reduce the computational cost of the algorithm.
c) Remove noise.
d) Make the results easy to understand.

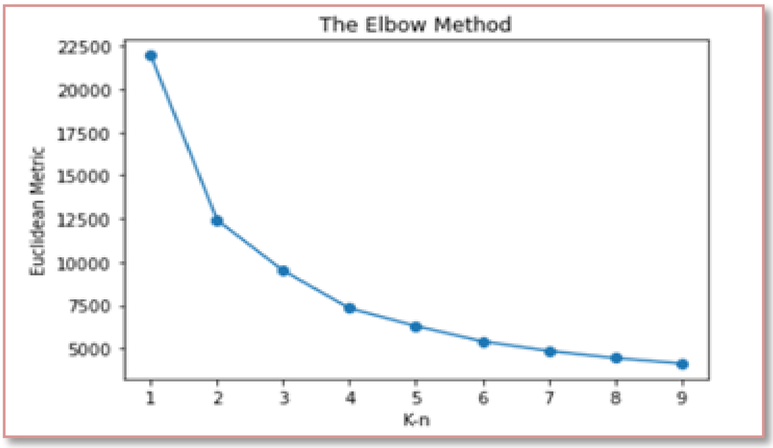
To make the following visualization and explanation of K-means much easier to be understood, we plan to reduce the dimension from 11 to 3. We also check the new 3 features explained variance ratio, this standard can show how much the contribution of new features is if we build a model with these new features. In another word, this indicator represents the data retention rate. Fortunately, our total explained variance ratio is 50%, we think this dimensionality reduction is effective. And we define these three new features as PCA-features.

1.2 K-means for clustering customer

The process of constructing user portraits is a kind of unsupervised machine learning. There are many algorithms that can promise this process to get a comfortable classification, of which K-means algorithm is a very effective method. The steps of K-means are as follows.

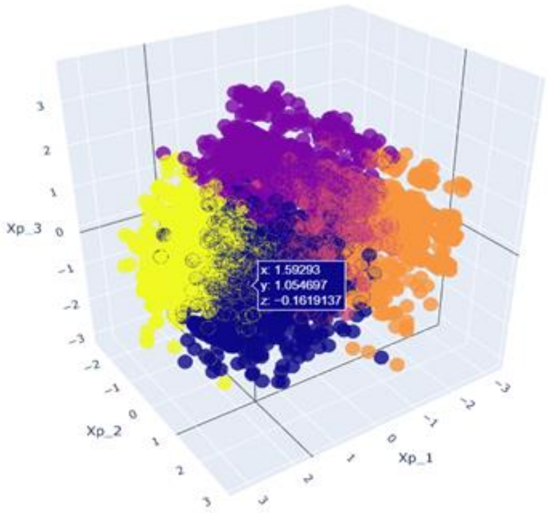
Step1: Finding suitable K by using the Elbow Method.

We calculate the Euclidean Metric of PCA-features and draw a line chart to visualize the result. According to the below chart, we can easily get K=5, the point is the inflection point when K=5, because the image tends to be flat.



Step2: K-means-process and visualization.

We will divide the customer into 5 clusters according to the 3 PCA-features and build a 3-dimension scatter plot for visual presentation. From the graph, we can roughly see the distribution of clusters.

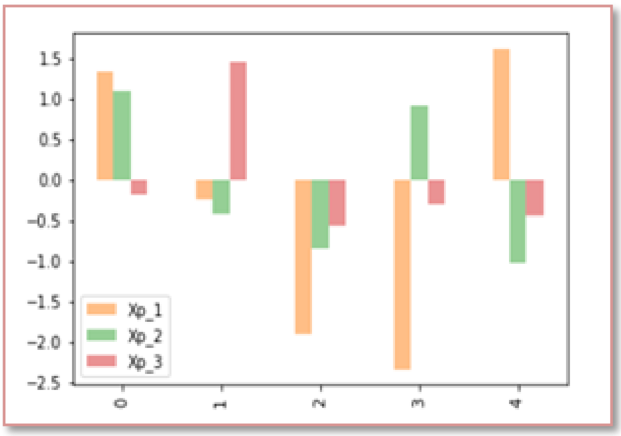


Step3: Evaluate the model

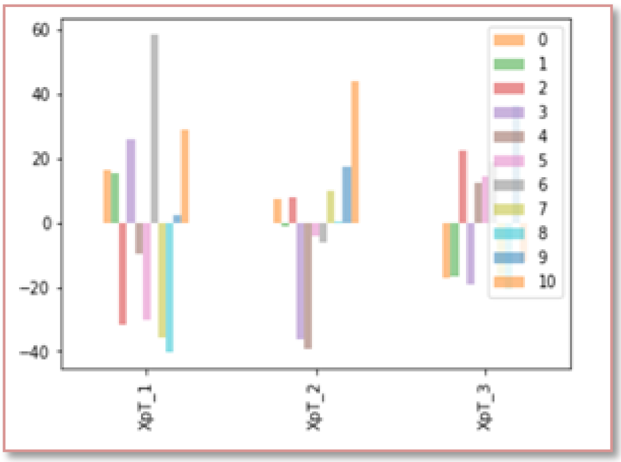
After the model is generated, the evaluation, quantifying the pros and cons of the model, is necessary. And the Silhouette Coefficient is generally used to evaluate the K-means model. The optimum value of Silhouette Coefficient is 1, the worst value is -1. The value close to 0 represents overlapping clusters. Negative values typically indicate that the sample has been assigned to the wrong cluster, because different clusters are more similar. But if the value is next to 1, it perhaps means that the model has risk of over-fitting. Fortunately, the Silhouette Coefficient of this model is 0.47, which means the model is good.

1.3 Build and analyze customer portrait

According to the K-means, we have already divided users into five categories. But the data is unmeaningful unless it is understood, so we implement properly Deconstructing the data and analyzing the background of data to get the readable explanation of the customer. We firstly deconstruct the center of each cluster into the metric of PCA-features as bar chart b1 and deconstruct the PCA-features into raw features as bar chart b2. So, we can see the raw features which mainly influence the K-means model.



Bar chart: b1



Bar chart: b2

Then we analyze the mean of each feature and the actual situation. We find that the feature named ‘CreditCardType’ ‘Gender’ distributes no specificity in different clusters, so we cannot use these 2 features to describe customer portraits. After analyzing the descriptive result of useful features – ‘'Age' 'Location' 'MotorValue' 'MotorType' 'HealthType' 'HealthDependentsAdults' 'HealthDependentsKids' 'TravelType' 'PrefChannel'. We finally create the following customer portrait. And the details of each customer portrait will be described in the marketing strategies report.

Predict communication channel

When users purchase ABC insurance products, the company provides three preferred communication channels, SMS, Phone and Email. We can use the accuracy of randomly selecting a channel as the baseline for selecting a channel, which is 33.3%. If the result of our predictive model is better than this baseline, we will use this predictive model to generate users' preference marketing channels.

Feature Engineering

Feature Selection

We did not select these features as our training data: CustomerID, GivenName, MiddleInitial, Surname and Occupation, because these features have no relationship with our product or have very high cardinality with low frequency.

Feature encoding

Categorical features are encoded by one hot encoding. Numeric features are encoded by minmaxscaler and yeo-johnson transformer, which normalizes features intnew range between 0 and 1 and more gaussian respectively.

During the modeling process, we used stratified cross validation (10 splits and 3 repeats) to test the possible weak estimators. The accuracy is as follow:

AdaBoostClassifier: 64.5%
BaggingClassifier: 60%
ExtraTreesClassifier: 58.3%
GradientBoostingClassifier: 65%
RandomForestClassifier: 62.2%
KNeighborsClassifier: 60%
SVC: 66.6%
XGBClassifier: 62.4%

After testing default setup and tuning the hyperparameter of possible algorithms, we selected five models as the weak estimators, SVC, AdaBoostClassifier, ExtraTreesClassifier (exported by TPOT), GradientBoostingClassifier and XGBClassifier respectively. We asked them to vote on the prediction results. The final prediction accuracy rate was 66% with 0.021 standard deviation.

Detailed evaluation report is as follows:

	precision	recall	f1-score	support
Email	0.63	0.70	0.66	5298
Phone	0.74	0.73	0.74	4689
SMS	0.56	0.44	0.49	2265

Overall Prediction Accuracy: 66.4% (standard deviation: 0.021).

Precision talks about how precise/accurate this model is out of those predicted positive, how many of them are actual positive. For example, the precision of the Phone means that 74% of preferred channels marked as Phone are true. This is a good measure when the cost of false positive is high. Considering calling customers is a strong disruptive behavior and it may adversely affect our brand if the phone is not customers' preferred channel, we use precision as the measure of accuracy.

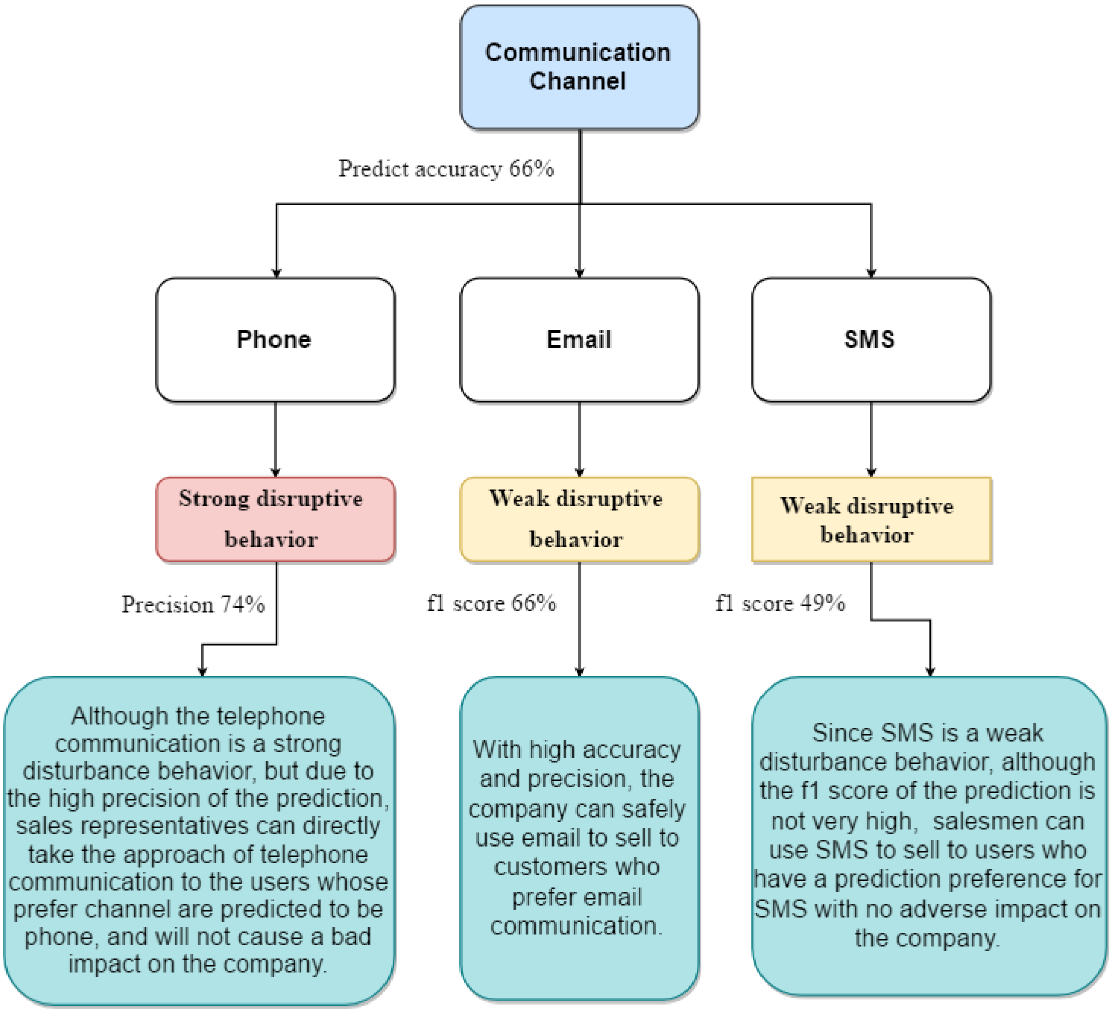
So Recall actually calculates how many of the Actual Positives our model captures through labeling it as Positive (True Positive). Since email and SMS will not disrupt customers much, we use f1 score (balance precision and recall) to measure their accuracy.

Final evaluation of the model as follow:

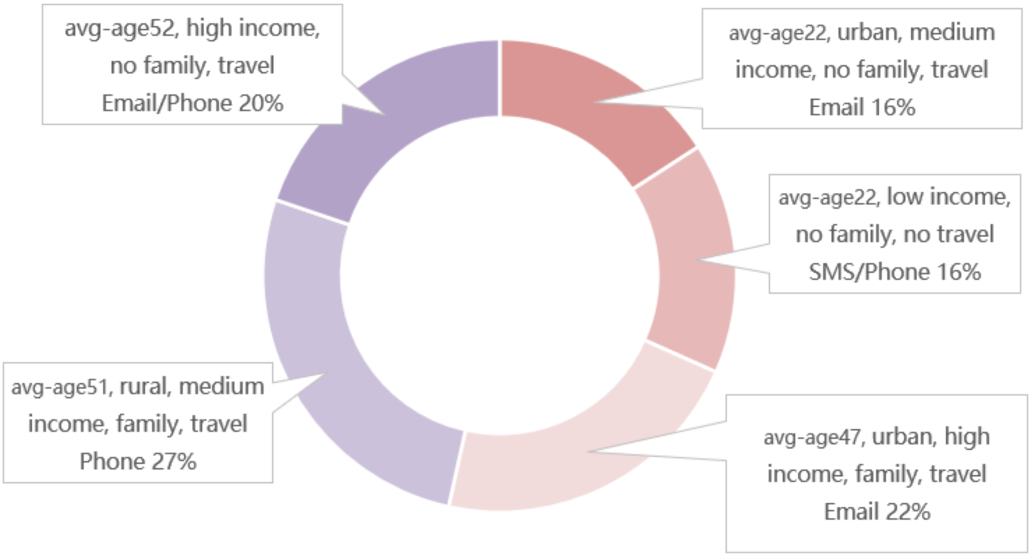
We use precision as the measure of predictive accuracy for Phone: 74%.
We use f1 score as the measure of predictive accuracy for Email and SMS, 66% and 49% respectively.

Marketing Strategy Report

As the model has been trained and built, after doing the same feature engineering on customer's data (customers without preferred channels), these data can be input into the model to predict the preferred channels. The final prediction data, feature engineering pipeline and model have been exported, which file names are Customer_data_with_PrefChannels.csv, transformer.pkl and model.pkl in the folder. Marketing team can use this file to check customers' data including their preferred channels.


Communication Channel
Predict accuracy 66%
Phone
Email
SMS
Strong disruptive behavior
Weak disruptive behavior
Weak disruptive behavior
Precision 74%
f1 score 66%
f1 score 49%
Although the telephone communication is a strong disturbance behavior, but due to the high precision of the prediction, sales representatives can directly take the approach of telephone communication to the users whose prefer channel are predicted to be phone, and will not cause a bad impact on the company.
With high accuracy and precision, the company can safely use email to sell to customers who prefer email communication.
Since SMS is a weak disturbance behavior, although the f1 score of the prediction is not very high, salesmen can use SMS to sell to users who have a prediction preference for SMS with no adverse impact on the company.

The analysis results show that the prediction results can be directly applied to sales work. As the customer portrait has already been constructed before, the marketing department can identify portrait portraits by adding user information, and then produce corresponding marketing strategies.



Users with an average age of 47 are the biggest potential users of home Insurance. First of all, they have several family members and their house is larger. The regular maintenance of the house and the safety of the facilities are very important to their family life. Secondly, high value motorcycles indicate that they have the characteristics of a high income and stable financial situation. Many of them are business leaders, whose high insurance costs are nothing compared to the security of their house and the stability of their families. It is no exaggeration to say that home insurance is an absolutely indispensable part of their insurance products list. It is also best to place advertisements through email channels on weekends to avoid being overlooked when messages are mixed with work content.

Users with an average age of 52 are another large group of potential customers. They have large families and choose to live in the rural area because their middle income may not be able to afford the high housing costs in cities. But the low price of rural homes means their houses are likely to be larger, so house maintenance and so on are essential.

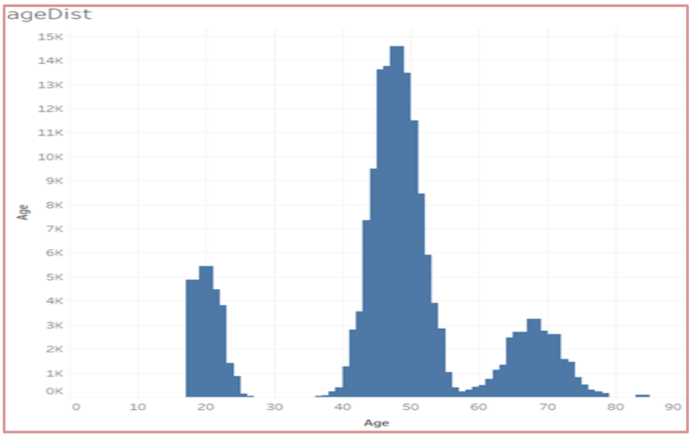
Young people with an average age of 22 and living in cities without a family are another group that will buy home insurance. They have a middle income and may have just bought a house in the city. For a person who does not have much energy to take care of the house and occasionally needs to travel, entrusting the house to the insurance product is the most ideal way. When a house encounters a problem, it can be solved effectively without having a significant adverse impact on its own financial situation.

Older age groups with an average age of 51, no family and high income are less likely to buy home insurance. They travel a lot and don't have much time to live at home. For various reasons they don't have a family and don't need a very big house to live in. They are only likely to buy home insurance if they own their house and rent it out to someone else, so that when they are not at home, tenants will be able to deal with the house problems under the insurance. Therefore, this group needs to check with them whether they need home insurance by phone or email.

The last group is the least likely to buy the insurance. They have low income, young age and no family. They may live in other people's houses. Therefore, it is not recommended to sell home insurance to these users.

In the meantime, there are some other suggestions.

First of all, it is interesting to note that the number of users aged between 30 and 40 is very small, only 9, and they are all distributed at the age of 38 or 39, which is close to 40. Why is there such a fault distribution? The company can conduct further mining and research on this part of the population, but the current data is not enough to explain the cause of this phenomenon.



Secondly, because the existing data set is too small, in the subsequent process of product promotion, users' preference channels can be asked to confirm the accuracy of the prediction, so as to better modify the prediction model and make the prediction results more reliable.

Reference

Trang T. Le, Weixuan Fu and Jason H. Moore (2020). Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics.36(1): 250-256.